

本周周报(12.24-1.6):

解聪

本周工作:

1. 论文书写

在原有的中文文档的基础上, 使用 latex 先写英文初稿。初步整理了论文的结构。

英文部分先写了 Introduction 以及 Related Work。相关工作这部分整理的还不是很完善, 主要是用户行为分析这一部分补充的不够。

数据分析的部分开始使用英文在写, 但是因为目前使用数据分析的方法不是很完善, 所以写的不是很完整, 只是初步介绍了数据。

2. 时序用户数据分析:

论文写作过程中我总结现有的工作, 现有问题在于如下几点:

1) 数据分析

淘宝数据分析的概率模型不够完善。但是目前还没有对高维的多用户数据的分析思路。

关于这个问题我请教了何晓飞那边的博士三年级学生胡尧。

我向他介绍交易记录数据的一些情况, 可以看做是高维时序数据。但它区别于普通高维时序的数据: 每时刻不仅仅只存在一组交易属性, 即一条交易记录, 而是存在许多的高维交易记录。

简单地说每个时刻对应多个高维向量, 而不是单个高维向量。并且在时刻 A 出现的特定 ID 的数据不一定出现在下一时刻。

问题是我们需要从交易数据中发现一些异常记录。他表示这个问题现有方法不能很好的解决。他之前做过一系列的类似时序数据的工作包括股票数据分析等, 但我们的问题更加复杂。

虽然现有工作都无法解决我们的问题, 他提出了两个近似的解决方案:

- 训练某一个模型, 针对每条记录分别判断异常与否。
- 寻找某种方法表示某一时刻整体数据的特征。比如将每个时刻的多个高维向量转化为单个向量, 保留多个向量的总体异常程度。按照时序检测的方法, 检测出某个时间段的异常, 再对这一时间段的所有高维向量检测。

第一种方法的问题比较大。一方面没有很好的模型判断异常, 这也是我们之前一直存在的问题; 另一方面对仅仅对单个向量判断会缺少上下文信息, 比如会忽略掉特定 ID 出现的频率信息。

第二种方法刚好符合我们的可视分析流程。它允许用户先选取一个特定的时间段, 再对该时间段进行可视分析。

虽然两个方法都不是很完美, 目前这一部分显然不需要做的很精确, 因为淘宝数据只是其中的一个案例而已, 这个工作还是把重点放在可视设计上。

Twitter 数据分析的文本挖掘算法空缺。

这个问题可以使用现有文本挖掘算法解决。包括主题抽取的方法, 如 LDA, TFIDF 等。微博情感分析, 目前尚不明确用什么方法。

2)可视编码。

视觉设计一直是工作的核心部分，但是目前的主要问题就是需要较多学习成本。

3)部分案例还很不完善。

Twitter 数据不够完整。目前的数据是 1600 多人的用户群的 Twitter 信息，解决问题的方法是重新抓一些有用的，可以反应时序模式的数据。

通讯数据不够具有说服力。目前从通讯数据中看出来时序特征不多，包括用户打电话模式的转换（本地->长途），其他信息则体现的比较少。一方面是因为数据中和时序有关的信息不足；另一方面可能通讯数据更适合做用户关系的分析。

另外本周修改部分结果。主要是对 Twitter 与通讯数据可视化的局部修改。但是这两个例子的数据对应的可视化效果一般，所以有必要选取更加适合的数据，比如按照我们的需求重新提取 Twitter 数据。

3. Twitter 数据的抓取

为了使 Twitter 数据的可视化更加完善，计划抓取适合的数据做方法的验证。

之前请教了汪飞，可以使用 opencalais 提供的服务获取所需要的 Twitter 数据，目前还在尝试中。

4. Taobao 的其他工作

我们在 Taobao 以后的工作不局限于已经有的几个项目，玄澄希望我们能够使用可视化帮助更多的产品。

目前，除了 DataVjs 以外。我们确定加入的有 TCIF 官网的可视化，这 and 用户标签的可视化研究相契合，可以利用一些现有的成果。另外也可以进一步明确用户标签可视化里的需求。另外还有其他一些项目，包括数据魔方产品和指数有一些可视化的需求。正在接触，还没有确定加入。

总的来说落地到数据产品中的可视化要求简单易懂，不需要很强的创新，工程性比较强一点。

下周工作：

1. Twitter 数据的抓取。
2. 时序用户行为可视化英文文档书写